HTML Tag Based Metrics for use in Web Page Type Classification

Jonathan Elsas

School of Information and Library Science, University of North Carolina at Chapel Hill, CB#3360, 100 Manning Hall, Chapel Hill, NC 27599-3306. Email: jelsas@email.unc.edu.

Miles Efron

School of Information and Library Science, University of North Carolina at Chapel Hill, CB#3360, 100 Manning Hall, Chapel Hill, NC 27599-3306. Email: efrom@unc.edu.

Traditional machine learning classifications of HTML documents focus on features drawn from terms in the documents, the link structure of groups of documents, or a combination of both. These techniques attempt to generate topical classifications of documents, with the hopes of mirroring a human's classification of pages into subject areas, thus facilitating retrieval. This paper presents an alternative method that aims at generating a "type-wise" classification of HTML documents. The types explored in this paper include tables, indexes, tables of contents, and textual content pages. These types of pages are of particular significance to the classification of documents on statistical web sites, which is one goal of the GovStat Project (http://www.ils.unc.edu/govstat), but also hold significance to HTML document collections at large.

Introduction

The extraction of specific nuggets of information from large bodies of hypertext has become a critical goal with the expansion of the World Wide Web and the public's growing reliance on this medium as a primary source of information. The WWW contains an incredible variety of documents, including press releases, web journals, retail catalogs, personal home pages, and large indexes. All these types of documents serve different purposes and it follows that most types of documents will not be relevant for any one information need. The development of specialized search and indexing techniques that are able to differentiate between different types of documents could be tailored to the information needs of a particular query, and therefore yield more targeted and superior results. This paper defines several document types and explores a technique for differentiating between those types in a governmental statistics web site.

The motivation for this work stems from several areas: a need to develop a mechanism to specify the document type in a retrieval situation, and to improve the performance of unsupervised and semi-supervised machine learning for large web sites. In the GovStat project (http://www.ils.unc.edu/govstat), we are interested in generating metadata by extracting the variable names and headings from statistical tables presented on government agencies' web sites. In order to do this, we must first retrieve pages that are likely to contain statistical data tables, as opposed to press releases, indexes, or other types of pages. The techniques presented in this paper should provide an efficient means of identifying a body of table-type pages, as well as other types.

In addition to this direct application to document retrieval, this classification scheme may also have more indirect IR applications. In large scale web sites, there are many pages that have only a navigational purpose: indexes, site maps, menus, etc. Because these pages often attempt to provide an overview of the topics in the web site, or a portion of the web site, they necessarily contain terms spanning much of the conceptual space of the entire corpus of documents. When these pages are used in large scale text-based clustering, they have a tendency to degrade clustering performance: terms that may not be topically related are associated through their co-occurrence in these index-type pages. By eliminating these kinds of pages from training sets, it is possible that automatic classification performance may improve. On the other hand, since these pages are often meant to be descriptive of all areas of the site, the use of link text from index pages may aid in labeling of clusters or augmenting document text throughout the site (Chakrabarti, 2000).

Related Research

Most of the IR research on hypertext classification in the WWW has been focused on primarily two areas: textual features of a document, or the topological structure of a body of hypertext documents. The textual features of hypertext documents are typically used in an analogous way to textual features in other document collections (Barfourosh et al., 2002). The text in the body of a document is often selectively augmented with text from within the <TITLE> or <META> tags, from the full text of pages that link to that page (Yang et al., 2002, Riboni, 2003), or from the anchor text of those pages (Chakrabarti, 2000).

Many metrics have been developed out of a body of hypertext's topological structure, and some of these metrics do define a type-based or quality-based classification. These classifications include Kleinberg's hub and authority types (Kleinberg, 1999), PageRank (Brin and Page, 1998) used by Google (http://www.google.com), and Botafogo et. al.'s (1992) index and reference type nodes. These do offer some degree of type-based classification, but none are based on the content of the document or attempt to identify tables.

Human identification of document types based on document structure has been explored by Toms, et. al. (1999). In their research, they found for specific document types, such as letters and journal articles, there is a strong link between identification of document types and the structure of the document. Although their research has not been extended into the development of automatic methods for document type classification, the strong correlation between document types and document structure suggests that development of machine-based methods are plausible.

Methodology

The analysis conducted here is intended to develop metrics that could be used to automatically classify HTML documents into several type-based categories. The scope of this classification is limited to three types of documents, but the techniques presented here could be generalized as other similar HTML document type classifications are identified. We will first define the metrics that will be used for our classification scheme, then define the types of documents to be identified, and explain the methodology for developing the thresholds used for document classification.

Metrics Definitions

In this section, several metrics for hypertext are presented. The calculation of these metrics requires considerable processing time due to the need for complete parsing of the DOM tree of each HTML document in a corpus, but the metrics can all be calculated as part of the indexing process and do not need to be computed in real-time.

Table Tag Ratio (*TTR*) This is the ratio of table tags to total tags in an HTML document. A table tag is defined to be one of the HTML elements: `<TABLE>`, `<TR>`, or `<TD>`. This metric could also be modified to include the `<THEAD>`, `<TBODY>` and `<TFOOT>` tags but these are used much less frequently and will not be considered for this study. This metric may also be simplified to include only the `<TD>` tag, as this tag should be an accurate indicator of how many table cells are in a document.

Anchor Text Ratio (*ATR*) This is the ratio of anchor text characters to the total text characters of an HTML document. The total text of a document is defined as the content of all DOM text nodes in the document, and the anchor text is the content of the DOM text nodes that are children of anchor (`<A>`) elements[1]. Anchor text in this study is the same as anchor text as referred to by Fürnkranz (1999) and Yang (2002).

Text Per Table Data Tag (*TTD*) This is the mean number of the text characters within a table data tag (`<TD>`) in the document. This is calculated by dividing the total number of characters in DOM text nodes that are children of table data (`<TD>`) tags by the total number of table data tags in a document. Of all the metrics presented, this is the only one that is not normalized to be between 0 and 1 – the total number of characters in a document bound this metric. It may be advantageous in some situations to normalize this metric to be between 0 and 1 also, and use an "undefined" value to represent documents that do not have any table tags. For the purposes of this study, however, this metric is left in its non-normalized form.

Web Page Type Definitions

The types of web pages we are interested in identifying with the above metrics are defined below:

(data) table A (data) table type page contains data arranged in rows and columns. This data typically represents comparative or time-series data. The cells in the table typically do not contain large amounts of natural language text, but rather contain numbers, single words, or short phrases. In the data set studied, large portions of the pages are of this type.

index/table of contents An index or table of contents page primarily contains hypertext links to other pages in a web site. These pages, also known as site maps, may be intended to be used as a starting point for browsing a site and typically serve only navigational purposes. These links are most often organized alphabetically or topically in a large list, and span most of the content areas of the web site. For the purposes of this study, we will treat index, site map, and table of contents pages interchangeably, although there are some obvious differences in the function of those types of pages. It may be useful to note that "index" pages here can be thought of as analogous to the graph-theoretic definition of an index node presented by Botafogo, et. al. (1992).

content A content page contains a significant portion of natural language or free text. For the purposes of this study, the content-type page category is a catchall category: all pages not classified as tables or indexes will be considered

[1] See http://www.w3.org/DOM for DOM related information and http://www.w3.org/TR/DOM-Level-2-HTML/ for details on the HTML DOM. Throughout this paper, we will refer to the HTML element names rather than the DOM Object names.

content pages. In applying traditional machine learning techniques to a body of HTML documents, reducing the set of document to these types of pages may result in better clustering performance.

These definitions are fairly subjective and are tied to unquantifiable aspects of the documents such as the intended use of the page. Using the metrics defined above, however, we can develop rigorous quantitative definitions of these page types.

Let *S* be the set of all web pages in a web site. A web site is understood to be all the HTML documents residing under a single domain name, accessible via any number of hops from a unique home page.

table Let the set of table pages, *T*, be the pages in *S* such that the *TTR* is greater than some threshold τ_{TTR} and *TTD* is less than some threshold τ_{TTD}. That is, the ratio of table tags to total tags is sufficiently high and the average number of characters per table tag is sufficiently low.

$$T=\{d \in S \text{ such that } TTR(d)>\tau_{TTR} \text{ and } TTD(d)<\tau_{TTD}\}$$

index Let the set of index pages, *I*, be the pages in *S* such that the *ATR* is greater than some threshold τ_{ATR}. That is, the number of characters in anchor text is a sufficiently high percentage of the total number of text characters in the document.

$$I=\{d \in S \text{ such that } ATR(d)>\tau_{ATR}\}$$

content For the purposes of this study, the set of content pages *C* is the remainder of pages in S.

$$C=S-(T \cup I)$$

The threshold values in the above definitions are highly dependent on the nature of the HTML documents being analyzed. For some bodies of HTML documents, there may not be clear qualitative divisions between the **index** and **content** type pages. Other bodies of HTML documents may not have any pages that would fall into the **table** category. For the large bodies of HTML pages analyzed for the GovStat project, however, the divisions between these types of pages are fairly clear.

Threshold Development

In order to develop the thresholds τ_{ATR}, τ_{TTR}, and τ_{TTD}, a sampling of documents was chosen from the Energy Information Administration's (EIA) web site[2] (http://www.eia.doe.gov) to be typical documents for each of the three categories. These samples are shown to be statistically distinct and then thresholds are developed based on these samples.

Sample Documents Used

The documents sampled from the EIA web site were chosen based on knowledge of the structure of the site using several simple heuristics. These heuristics include the document's location in the file structure of the web site, file-naming conventions used in parts of the web site, and conventions used in the document titles. These rules were chosen because they presented a simple and convenient means to identify documents of the specific types. Although these pages are chosen to be exemplars of the pages in each category, it is possible (and in fact likely) that there may be some pages that are outliers. But, of the body of documents chosen for each category, it is expected that the vast majority are in fact correct examples of that type.

SampleTablePages 736 pages were chosen as table type pages. These page have the string "Table" contained in the HTML `<TITLE>` tag, have the string "tbl" in their filename, and contain at least one table tag (see the definition of *TTR*, above).

Sample Index Pages 56 pages were chosen as index type pages. These pages were chosen from various areas of the web site, including the text-based index pages, archived press release indexes, and departmental directory indexes[3].

Sample Content Pages 174 Content type pages were chosen based on their file name containing the string "press" in the file and located in the "Press Release" directory on the web site (/neic/press/). These press release pages have a tendency to have a large amount of textual content, a small number of anchors and do not generally display large amounts of tabular data.

Analysis of Samples

Evaluating the means of the metrics for the sample types of documents shows some distinct differences between the types of pages (Table 1). In particular, note that the **table** type pages have the highest mean *TTR*, and the **index** type pages have the highest mean *ATR*.

In order to show that the sample pages are statistically distinct sets of pages with regard to the metrics, a simple T-Test was performed on the data (Table 2). Based on these results, we can see that there is a minimal probability that the three groups of samples came from the same distribution when looking at all three of these metrics. It

[2] Documents from the Energy Information Administration (EIA) web site, http://www.eia.doe.gov, were downloaded on December 12, 2003. All analysis is based on the documents present on the EIA site at that time.

[3] The regular expressions used to match the index page URLs are: .*njava.*\.htm.*, .*/neic/months.*\.htm., and .*/bookshelf/InfoDir2001/.*index.*\.htm.*.

is expected that the table and index types have similar distributions for the *TTD* metric since table tags are often used for the organization of links in an index-type page. It is also not surprising that the table and content types have similar distributions for the *ATR* values because this metrics will primarily be used for distinguishing the index-type pages.

Table 1: Means and Standard Deviations of the metrics for each type.

Type		***TTD***	***ATR***	***TTR***
Table	Mean	31.675	0.033	0.357
	st. dev.	437.830	0.036	0.103
Index	Mean	30.351	0.642	0.233
	st. dev.	12.537	0.103	0.073
Content	Mean	349.605	0.032	0.043
	st. dev.	385.670	0.018	0.055

Table 2: *p* values from T-Tests for each pair-wise combination of types.

Type 1	**Type 2**	***TTD***	***ATR***	***TTR***
Table	Index	0.468	0.000	0.000
Table	Content	0.000	0.295	0.000
Index	Content	0.000	0.000	0.000

Thresholds

In order to develop the actual threshold values based on the sample documents, we can simply take the midpoint between the means of the category in question and the remaining categories. For example, since the *TTR* value is an identifier for the table-type pages, we can assign the threshold value, τ_{TTR}, to midpoint between the mean *TTR* value for table pages, and the mean *TTR* value for non-table pages (0.089). The calculation of the threshold is given below:

$$\begin{aligned}\tau_{TTR} &= (\mu_{TTR}(\text{tables})+\mu_{TTR}(!\ \text{tables}))/2 \\ &= (0.357+0.089)/2 \\ &= 0.223\end{aligned}$$

where μ_{TTR}(tables) is the mean *TTR* value for tables and μ_{TTR}(! tables) is the mean *TTR* value for non-tables. Analogously, we can calculate the other threshold values, noting that *TTD* is another identifier for the table-type pages and *ATR* is the identifier for the index-type pages.

$$\begin{aligned}\tau_{TTD} &= (\mu_{TTD}(\text{tables})+\mu_{TTD}(!\ \text{tables}))/2 \\ &= 151.78 \\ \tau_{ATR} &= (\mu_{ATR}(\text{indexes})+\mu_{ATR}(!\ \text{indexes}))/2 \\ &= 0.337\end{aligned}$$

These three threshold values give us hard cutoffs for defining the three types of documents.

Other approaches to developing these thresholds will be considered for future work. These approaches could include clustering the documents into three classes using *k*-means, developing a Support Vector Machine based model for the classification, or other methods that would yield a more flexible and complex division between the classes (Mitchell, 1997). The method outlined above, provides a fairly accurate classification scheme with regards to the sample documents, and was considered sufficient for this study.

Results

We can now evaluate the thresholds against the sample training documents, and also classify the documents in the entire site. When looking at the body of training documents, and evaluating them using the rules defined in the "Web Page Types" section, we can see that the thresholds are about 95% accurate. 699 tables (94.97%) classified correctly as tables, and 55 indexes (98.21%) classified correctly as indexes. Further analysis is given in Table 3. Note that, based on the above definitions, the categories **table** and **index** are not mutually exclusive and this is reflected in classification of the index documents. Although almost half of the indexes classified as table pages also, it is advisable that the index categorization take precedence, and these pages not be treated as tables. That is, when classifying the pages in an entire web site, first classify the index pages, and then classify the remaining pages as tables or content.

Table 3: Analysis of the Training Set

		Classified as category			
Documents	**Total**	**Table**	**Index**	**Content**	**Accuracy**
Table	736	699	0	37	94.97%
Index	56	25	55	1	98.21%
Content	174	3	0	171	98.28%

Using the same threshold values, we can now look at the entire web site and determine what portion of the pages are table, index, or content types. Of the approximately twenty thousand HTML pages evaluated from the EIA web site, almost 60% classified as tables, 6.5% as indexes and 36% as content pages (see Table 4).

Table 4: Analysis of the Entire Web Site

Page Type	Number	Percentage of Site
Index	1306	6.46%
Table	11607	57.44%
Content	7294	36.10%
Total Pages	20207	

Conclusions and Future Work

The metrics presented here offer a novel way to distinguish HTML documents that are more index-like or more table-like. The results presented cover just a single web site, but it is clear from these results that overall trends are likely to develop across other sites. Future studies applying these metrics to other web sites should provide insight into development of general rules for choosing the metrics' threshold values.

An initial motive for this research was the hypothesis that removing the index-type and table-type page from the training set of a text-based classifier would improve the classification performance. This hypothesis has not been tested and is an obvious direction for future research. Another unexplored future research direction includes using these metrics as features for traditional machine learning techniques in order to automatically cluster a body of HTML documents into the defined types instead of explicitly deriving the threshold values. Due to the small number of metrics defined here, and thus the small feature set to use in clustering, machine-learning algorithms should perform well.

The primary risk in using these metrics, and any metrics based on HTML tags, is that HTML code is not standardized, is subject to the quirks of HTML authors and authoring tools and can be quite stylistic. Many HTML authors typically use table tags to facilitate layout, and the presence of those tags does not imply the presence of a table as defined above. The definition of the *TTD* metric is an attempt to mitigate these concerns, however it is possible that more complicated layout schemes could affect the classifications. Future research should provide insight into how effective this metric is in separating complicated table layout from true data tables, and how sensitive the classification is to the quirks of different authoring tools.

Finally, these metrics are based on fairly crude analyses of the documents: raw counts of tags and characters in HTML that don't take into account how those tags are being used. This can be thought of analogous to the existing, equally crude, text based features commonly used in IR, such as TF-IDF (Chakrabarti, 2000). Because of the smaller set of "terms" in HTML and HTML's relative lack of complexity compared to natural language, it may be possible to develop more robust and intelligent metrics that could be used for the type of classification presented in this paper. Through the research for this paper, it is clear that the use of HTML tags in hypertext documents is quite rich and nuanced, and much more can be learned from analyzing the use of these tags.

ACKNOWLEDGMENTS

This research was made possible by the generous support of the National Science Foundation (NSF) under Grant NSF EIA 0131824. The author would also like to thank Gary Marchionini and Stephanie Haas for their support and guidance.

REFERENCES

Barfourosh, A. A., Nezhad, H. M., Anderson, M. L., and Perlis, D. (2002). Information retrieval on the world wide web and active logic: A survey and problem definition.

Botafogo, R. A., Rivlin, E., and Shneiderman, B. (1992). Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Trans. Inf. Syst.*, 10(2): 142–180.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117. Elsevier Science Publishers B. V.

Chakrabarti, S. (2000). Data mining for hypertext: a tutorial survey. *SIGKDD Explor. Newsl.*, 1(2): 1–11.

Fürnkranz, J. (1999). Exploiting structural information for text classification on the www. In *Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis*, pages 487–498. Springer-Verlag.

Kleinberg, J. M. (1999). Hubs, authorities, and communities. *ACM Comput. Surv.*, 31(4es): 5.

Mitchell, T. (1997). *Machine Learning*. Boston: McGraw Hill.

Riboni, D. (2003). Feature selection for web page classification. In A Min Tjoa, A. C. S., editor, *EURASIA-ICT 2002 Proceedings of the Workshops*.

Toms, E. G., Campbell, D. G., and Blades, R. (1999). Does genre define the shape of information? the role of form and function in user interaction with digital documents. In *Proceedings of the 62nd Annual Meeting of the American Society for Information Science*, pages 693–704. Information Today, Inc.

Yang, Y., Slattery, S., and Ghani, R. (2002). A study of approaches to hypertext categorization. *J. Intell. Inf. Syst.*, 18(2-3): 219–241.